Do SLM agents screen papers as well as LLMs?

João Goulart Mendes de Freitas Filho[1], **Sílvio R. Fernandes**[1]

[1]Departamento de Computação – Lab. GESyCA
Universidade Federal Rural do Semi-Árido (UFERSA)
Mossoró – RN – Brasil.

joao.filho63204@alunos.ufersa.edu.br, silvio@ufersa.edu.br

Abstract. *Summarizing scientific knowledge is crucial, and the Systematic Literature Review is one of the main methods used, especially in evidence-based research, being a complex and time-consuming process. This work focuses on the article screening stage, which is one of the most critical steps, as the subsequent phases depend on its quality. The literature shows promising results using commercial LLMs such as ChatGPT and Gemini for this task, but there is a lack of studies on the use of smaller language models (SLMs) running locally. We analyzed three approaches using SLMs in comparison to a commercial LLM used as a baseline. The results show good performance from SLMs when tasks are stratified and simplified into subtasks. Qwen 3 - 8B achieved an accuracy of up to 94.35%. Using a multi-agent approach, Phi 4 - 14B reached 79.5%, and Qwen 3 - 4B reached 78.8%, compared to the commercial LLM.*

Resumo. *Sumarizar o conhecimento científico é crucial, e a Revisão Sistemática da Literatura é um dos principais métodos utilizados, sobretudo na pesquisa baseada em evidências, sendo um processo complexo e demorado. Este trabalho foca na etapa triagem de artigos, sendo uma das mais cruciais, pois as demais dependem da qualidade desta. A literatura aponta bons resultados com LLMs comerciais como ChatGPT e Gemini para esta tarefa, mas faltam estudos sobre o uso de modelos menores (SLMs) executados localmente. Analisamos três abordagens com SLMs em comparação a um LLM comercial, usado como baseline. Os resultados mostram bom desempenho dos SLMs quando estratificamos e simplificamos em subtarefas. O Qwen 3 - 8B alcançou acurácia de até 94,35%. Com abordagem multiagente, o Phi 4 - 14B obteve 79,5% e o Qwen 3 - 4B, 78,8%, em relação ao LLM comercial.*

1. Introdução

A Revisão Sistemática da Literatura (RSL) é metodologicamente fundamental para a condução de pesquisas baseadas em evidências [Mulrow 1994]. A definição do estado da arte de um determinado tópico a partir da síntese rigorosa, sistemática e transparente das informações é imprescindível para direcionar e otimizar a pesquisa científica.A RSL é um processo inerentemente complexo e que demanda muitos recursos humanos e de tempo [Borah et al. 2017]. Sua execução é composta por fases bem definidas, conforme as diretrizes em [Kitchenham and Charters 2007, Page et al. 2021]. Geralmente, essas fases compreendem: (i) o planejamento a partir das questões de pesquisa e a formulação do protocolo a ser seguido; (ii) a condução da revisão que é composta pela busca dos estudos, a remoção de duplicatas, a triagem dos artigos encontrados, a extração dos dados dos

estudos selecionados, a avaliação da qualidade de cada trabalho; (iii) e, por fim, a síntese e apresentação dos achados.

Este trabalho tem como foco a fase de triagem dos artigos encontrados nas bases de dados a partir da análise de seus títulos e resumos, sendo uma das principais etapas e também um de seus principais gargalos. Em média, 20% do tempo total da revisão é investido nesta etapa [Haddaway and Westgate 2018], além disso, estudos indicam que, entre 80% e 97% dos trabalhos retornados das bases de dados são descartados [Sampson et al. 2011, Gusenbauer and Haddaway 2020]. Isso é uma consequência dos mecanismos de busca das bases onde os estudos são encontrados, que, mesmo incluindo a lógica *booleana* das palavras-chave, não aplicam contextualização semântica. Triar artigos trata-se de uma atividade trabalhosa e, aparentemente, mecanicamente repetitiva. No entanto, o fato de um pesquisador trabalhar em ciclos de análise, a aplicação dos critérios de inclusão/exclusão não é "determinista", exigindo uma ação cognitivamente semântica. Porém, isso também torna o processo suscetível a vieses e subjetividades por parte dos revisores. É neste ponto que os *Large Language Models* (*LLM*) podem auxiliar no processo de triagem, que precisa levar em consideração a semântica da aplicação de tais critérios sob diferentes formas de escritas e argumentações.

Segundo [Galli et al. 2025], os *LLMs* são candidatos em potencial para a otimização dessa etapa da RSL, pois apresentam capacidade de classificação *zero-shot/few-shot*, diminuindo drasticamente o esforço inicial em comparação a modelos tradicionais. *LLMs* têm fácil adaptação por meio de engenharia de *prompt*, ou seja, ajustes simples nas instruções e o modelo se adapta a outro cenário. O estudo também indica que o poder dos *LLMs* deve ser integrado às verificações humanas para que a integridade metodológica seja mantida e a alta performance seja garantida.

Segundo [Maslej et al. 2025], em 2022, o *LLM* com menor número de parâmetros que conseguiu obter uma pontuação acima de 60% no *benchmark MMLU* definido por [Hendrycks et al. 2021] foi o *PaLM* com 540 bilhões de parâmetros. Em 2024, o Phi-3 mini conseguiu o mesmo feito com apenas 3,8 bilhões de parâmetros, demonstrando um avanço notável e evidenciando a aplicabilidade de modelos menores em tarefas complexas. A quantidade de parâmetros destes modelos normalmente está na ordem de milhões ou unidades de bilhões, e por isso também são chamados de *Small Language Models - SLMs*, em contraste aos LLMs com centenas de bilhões ou até trilhões de parâmetros. Existe uma relação direta entre a quantidade de parâmetros e o desempenho em tarefas complexas [Kaplan et al. 2020], ao custo de muitos recursos computacionais. No entanto, [Hoffmann et al. 2022] demonstraram que modelos menores treinados com mais dados podem superar LLMs maiores subtreinados. Outros estudos indicam que SLMs podem aumentar significativamente seu desempenho com aplicação de técnicas de engenharia de *prompt* e capacidades agênticas, com as vantagens de baixa latência e menor custo [Luo et al. 2023, Belcak et al. 2025].

Portanto, o objetivo deste trabalho é investigar o uso de SLMs na triagem de artigos de uma RSL em comparação com um LLM comercial. Este artigo está organizado assim: na Seção 2 são apresentados os trabalhos relacionados; na Seção 3, a metodologia proposta; na Seção 4, os resultados obtidos e as discussões; e, por fim, na Seção 5, as conclusões e pesquisas futuras.

2. Trabalhos Relacionados

A IA e os *LLMs* estão mudando a investigação científica, impactando o modo como se conduz e valida a pesquisa [Mechanism 2024, Telenti et al. 2024]. Nesse contexto, a aplicação de modelos de linguagem na condução de RSL tornou-se uma área de grande interesse, como aponta [Galli et al. 2025], em especial na etapa de triagem de artigos (*screening*).

Estudos como [Delgado-Chaves et al. 2025] discutem o papel emergente dos *LLMs* na transformação dessa tarefa, este trabalho analisou o desempenho de 18 *LLMs* diferentes em três revisões sistemáticas já existentes, comparando suas classificações com revisores humanos. Os resultados indicam que os *LLMs* podem reduzir significativamente o esforço manual, além disso, o desempenho melhora quando os critérios de inclusão e exclusão são bem definidos. Investigações comparativas, como [Colangelo et al. 2025, Li et al. 2024], avaliaram a eficácia de diferentes *LLMs*, em geral modelos comerciais de grande escala, na tarefa de triagem. Estes estudos apontam redução de tempo e de esforço manual repetitivo em relação a revisões totalmente humanas.

Entretanto, uma lacuna ainda persiste: a aplicação de *SLMs* abertos nas etapas de RSL, sobretudo a triagem. Abordagens assim oferecem vantagens em custo, privacidade e acessibilidade. Além disso, também é interessante investigar múltiplas abordagens, avaliações de diferentes *prompts* e encadeamento de agentes baseados em *SLMs*, verificando se a divisão da tarefa em subtarefas mais simples influencia de alguma maneira o desempenho do sistema de triagem como um todo.

3. Metodologia

Este trabalho utilizou vários *SLMs* abertos por meio de três abordagens distintas para compará-los com um grande modelo comercial. Todos os modelos utilizados nos experimentos foram obtidos e executados diretamente pelo *Ollama* [Ollama Team and Contributors 2025] em suas versões quantizadas em 4 *bits* e configurados para executar com temperatura de 0,1. O *LangChain* [LangChain Team and Contributors 2025] foi utilizado como *framework* orquestrador dos *SLMs*, seus respectivos *prompts* e a interação entre eles. A Tabela 1 resume as características dos modelos utilizados. Além dos nomes/versões dos modelos, são apresentadas o tamanho (quantidade de bilhões de parâmetros) e se foram concebidos pelos desenvolvedores com capacidade de raciocínio, que neste contexto refere-se à habilidade do modelo em decompor problemas em etapas lógicas, como demonstrado pela técnica *Chain-of-Thoug*ht (Cadeia de Pensamento) [Shojaee et al. 2025]. A partir do tamanho, categorizamos os *SLMs* em ”Pequeno”(3B e 4B), ”Médio”(7B e 8B) e ”Grande”(12B e 14B).

A experimentação foi baseada em uma RSL tradicional que está sendo conduzida pelo grupo de pesquisa sobre geração de dados sintéticos para o reconhecimento de atividade humana. O protocolo desta RSL foi elaborado com os seguintes critérios de inclusão: *1. O trabalho propõe síntese de dados*; *2. O trabalho utiliza redes neurais generativas para gerar os dados*; *3. O trabalho aborda Reconhecimento de Atividade Humana* e a seguinte *string* de busca: *(”synthetic data”OR ”data augmentation”) AND (”human activity recognition”) AND (”smartphone”OR ”sensor”).*

Tabela 1. *SLMs* utilizados nas três abordagens

Nome e versão	Quantidade de Parâmetros	Categoria *SLM*	Raciocínio
Qwen 2.5	3B	Pequeno	Não
Phi 4 mini reasoning	3.8B	Pequeno	Sim
Gemma 3	4B	Pequeno	Não
Qwen 3	4B	Pequeno	Sim
DeepSeek R1	7B	Médio	Sim
Mistral v0.3	7B	Médio	Não
Qwen 2.5	7B	Médio	Não
Granite 3.3	8B	Médio	Não
LLaMA 3.1	8B	Médio	Não
Qwen 3	8B	Médio	Sim
Gemma 3	12B	Grande	Não
DeepSeek R1	14B	Grande	Sim
Phi-4	14B	Grande	Não
Qwen 2.5	14B	Grande	Não

Esta *string* foi adaptada para as bases de dados *Springer*, *IEEE*, *ACM* e *Elsevier*, com período de publicação entre janeiro de 2023 e março de 2025. Após remoção de duplicatas, restaram 619 trabalhos para serem triados a partir de seus títulos, *abstracts* e palavras-chave. Uma triagem humana foi realizada a partir dos critérios anteriores para a classificação de cada trabalho da seguinte forma: *INCLUÍDO* quando o trabalho atende a todos os 3 critérios definidos; *EXCLUÍDO 1* quando não atende ao primeiro critério; *EXCLUÍDO 2* quando não atende ao segundo critério e *EXCLUÍDO 3* quando não atende ao terceiro critério.

Em seguida, os mesmos trabalhos foram submetidos ao LLM comercial *Gemini 2.0 Flash Thinking Experimental 01-21* com instruções para classificar cada artigo como (*Included, Excluded 1, Excluded 2, Excluded 3* ou *Undeterminated*) com base nos títulos, *abstracts*, palavras-chave e os critérios de inclusão. Vale destacar que o modelo LLM pode classificar um trabalho como *Undeterminated* caso não consiga determinar se o enquadra nas demais classes. Além da classe, os modelos também foram instruídos a incluir uma justificativa para sua classificação.

As classificações dos artigos por este modelo foram comparadas às classificações humanas e coincidiram em 98%, apresentando fortes indícios de que o modelo segue bem as instruções para este assunto em particular. Vale destacar que este modelo classificou apenas 1 artigo como ”*undetermined*”, o que demonstra que ele confia em suas decisões. O *Gemini*, modelo utilizado como *baseline* retornou a seguinte classificação: 58 artigos marcados como *INCLUDED*, 445 marcados como *EXCLUDED 1*, 109 como *EXCLUDED 2*, 6 como *EXCLUDED 3* e apenas 1 como *UNDETERMINED*

Na **Abordagem 1**, foi utilizado um agente único que recebeu como entrada o contexto inicial, o direcionamento de sua tarefa, todos os três critérios de inclusão, título, resumo, palavras-chave e instruções para classificação para apenas 2 classes: *include* e *excluded* e dar uma breve justificativa de sua escolha, como mostra a Listagem 1.

Listagem 1. Prompt utilizado na *Abordagem 1*

```
You are an expert in article screening and should follow the inclusion criteria described below:
The inclusion criteria are:
1. The article proposes data synthesis;
2. The article uses generative neural networks to generate synthetic data;
3. The article addresses the recognition of human activities, regardless of whether the term "Human Activity
    ↪ Recognition" is explicitly used.
Your goal is to read the title and abstract of an article, and then provide a response indicating the classification
    ↪ ('INCLUDED' or 'EXCLUDED') and a brief justification for your classification. The article under review should
    ↪  be classified as INCLUDED only if ALL inclusion criteria are met, otherwise it should be classified as
    ↪ EXCLUDED.
Your output should be a JSON object with two keys: "classification" and "justification".
```

```
For example:
{"classification": "INCLUDED", "justification": "This article meets all three inclusion criteria."}
or
{"classification": "EXCLUDED", "justification": "This article does not meet inclusion criterion 1 as it does not
    ↪ propose data synthesis."}
      Title: {title}
      Abstract: {abstract}
```

Na **Abordagem 2** foi utilizado um agente único que recebeu as mesmas entradas com instruções para as 5 classes: *included*, *excluded 1*, *excluded 2*, *excluded 3* e *undetermined* e uma breve justificativa, conforme a Listagem 2.

Listagem 2. Prompt utilizado na *Abordagem 2*

```
You are an expert in article screening and should follow the inclusion criteria described below:
The inclusion criteria are:
    1. The article proposes data synthesis;
    2. The article uses generative neural networks to generate synthetic data;
    3. The article addresses the recognition of human activities, regardless of whether the term "Human Activity
        ↪ Recognition" is explicitly used.
Your goal is to read the title and abstract of an article and then provide a response indicating the classification ('
    ↪ INCLUDED', 'EXCLUDED 1', 'EXCLUDED 2', 'EXCLUDED 3' or 'UNDETERMINED') and a brief justification for your
    ↪ classification. The article under review should be classified as 'INCLUDED' only if ALL inclusion criteria
    ↪ are met; otherwise, it should be classified as 'EXCLUDED 1' if it does not meet inclusion criterion 1, '
    ↪ EXCLUDED 2' if it does not meet inclusion criterion 2 and 'EXCLUDED 3' if it does not meet inclusion
    ↪ criterion 3. If you cannot classify it in any of the above, classify it as 'UNDETERMINED'.
Your output should be a JSON object with two keys: "classification" and "justification".
For example:
{{"classification": "INCLUDED", "justification": "This article meets all three inclusion criteria because it proposes
    ↪ data synthesis using generative neural networks for human activity recognition."}}
or
{{"classification": "EXCLUDED 1", "justification": "This article does not meet inclusion criterion 1 as it does not
    ↪ propose data synthesis."}}
or
{{"classification": "UNDETERMINED", "justification": "For this article I cannot determine whether it can be included
    ↪ or does not meet one of the criteria for exclusion."}}

Title: {title}
Abstract: {abstract}
```

A **Abordagem 3** trata-se de uma solução multiagentes. Três agentes são direcionados a analisar cada um apenas um dos critérios de inclusão, retornando *True* ou *False*, e um quarto agente faz a classificação final seguindo uma lógica *booleana* que combina as respostas dos agentes anteriores de acordo com as seguintes regras: *Included* quando todos os agentes anteriores classificaram como *Included*; *Excluded* n onde n corresponde ao primeiro agente, na ordem de 1 a 3, que classificou como *False* e *Undetermined* se algum dos 3 agentes classificou como *Undetermined*. A Listagem 3 mostra o *prompt* utilizado nos 3 primeiros agentes, mudando apenas o critério de inclusão.

Listagem 3. Prompt utilizado nos *Agentes 1, 2 e 3*

```
You are an expert in screening scientific articles for literature review on the topic: Human Activity Recognition (HAR
    ↪ ).
You must follow the following criteria to indicate whether it should be included in a literature review.
Inclusion Criteria:
{inclusion_criteria}
Your task is to classify the article strictly according to the above criteria.
- If the article perfectly meets the inclusion criteria, classify it as "TRUE".
- If you are certain that the article does not meet the inclusion criteria, classify it as "FALSE".
- If there is insufficient information or you have doubts whether the article meets the inclusion criteria, classify
    ↪ it as "UNDETERMINATED".
Be clear and decisive. Avoid "UNDETERMINATED" unless absolutely necessary.
Your final answer should be ONLY a JSON object with the keys: "classification" and "justification".
Example output for you to follow:
{"classification": "TRUE", "justification": "The article proposes an approach to data synthesis."}
{"classification": "FALSE", "justification": "The article uses synthetic data but does not propose its synthesis."}
{"classification": "UNDETERMINATED", "justification": "There is not enough information in the title and abstract to
    ↪ determine."}
Now analyze the title and abstract of the following article and classify it.
Title: {title}
Abstract: {abstract}
```

4. Resultados

Para avaliar o desempenho dos *SLMs* na tarefa de triagem, as informações dos 619 artigos foram enviadas individualmente para cada um dos modelos em cada uma das 3 aborda-

gens. Em seguida, cada classificação dos *SLM* foi comparada à do *LLM baseline*.

É importante ressaltar que a classe *included* é especialmente importante para a RSL, não incluir trabalhos que deveriam (falsos negativos) é muito mais grave do que incluir aqueles que podiam ser descartados (falsos positivos), pois os falsos negativos não terão a chance de serem revisados por completo e isso pode impactar diretamente a qualidade da revisão. Os falsos positivos implicam em mais trabalhos para serem revisados por completo nas próximas fases, apesar de serem menos graves, muitos falsos positivos não são interessantes, pois a ferramenta perde o seu propósito de facilitar o trabalho humano.

4.1. Abordagem 1

Nesta abordagem, as classes possíveis eram: *"excluded"*, *"included"* e *"undetermined"*, com o intuito de avaliar principalmente a capacidade de incluir artigos corretamente, de modo que o motivo da exclusão não importasse. Dessa forma, todas as 3 decisões relativas à exclusão pelo *Gemini* foram condensadas para apenas *"excluded"*.

A Figura 1 apresenta a acurácia dos *SLMs*, ou seja, o percentual de concordância com as respostas do *Gemini* (barra mais à esquerda no gráfico) seguida pelos SLM dos menores até os maiores. Em geral, os modelos maiores tiveram melhor performance, entretanto, nas 3 categorias de tamanhos, houve modelos com pelo menos 90%. Os principais destaques são para o *qwen3_8b* com 94,35%, o *qwen2.5_7b* com 93,7% e o phi4_14b com 93,38%. Com exceção do qwen2.5_7b, os demais são modelos com raciocínio (LRM).

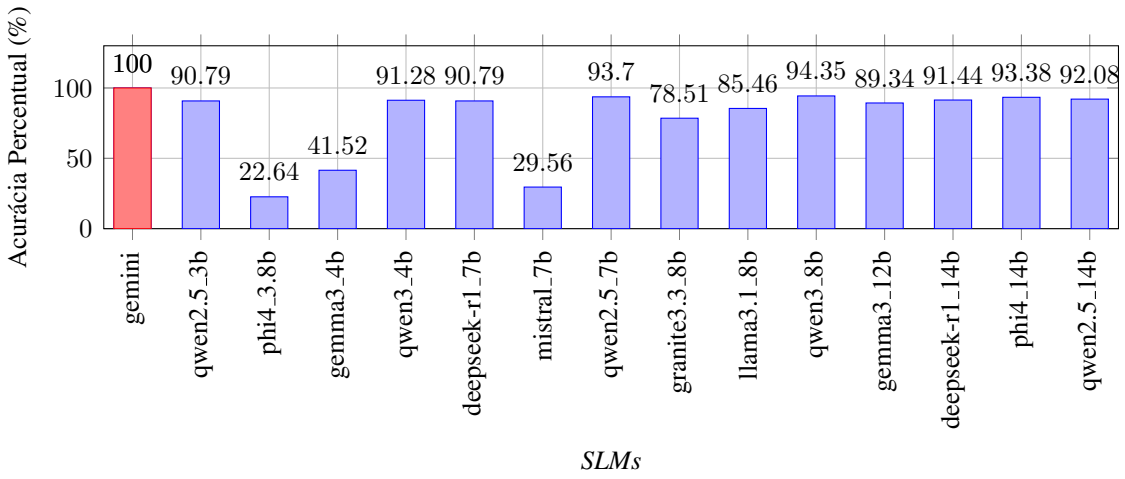

Figura 1. Percentual de concordância de cada *SLM* em relação ao *Gemini*

Apesar da *Abordagem 1* apresentar boa acurácia em relação ao *LLM baseline*, a rastreabilidade das decisões é essencial no processo de RSL, isso a torna insuficiente para esse tipo de tarefa, pois não define o motivo exato da exclusão e, portanto, não mantém a integridade metodológica do processo.

4.2. Abordagem 2

Os modelos *LRM* (Language Reasoning Model) *Qwen3:4b* e *Qwen3:8b* apresentaram problemas com o *prompt* utilizado nesta abordagem, esses *SLMs* entram em *looping* de

reasoning e não fornecem sua decisão em um tempo viável. Além disso, quando obtida, a saída é sem sentido e claramente um *looping* textual. Portanto, os gráficos desta abordagem não apresentam registros desses dois modelos.

A Figura 2 apresenta os quantitativos de cada *SLM* para as 5 classes adotadas nesta abordagem, incluindo as estratificações dos motivos dos excluídos. Neste caso, os modelos, em geral, classificaram muitos mais como *include*, especialmente os modelos *Gemma 3_4b* e *mistral_7b* , com um grande exagero. Ao se aumentar o número de classes, o percentual de concordância com o *baseline* (acurácia) decai em comparação a ***Abordagem 1***, como se vê na Figura 3.

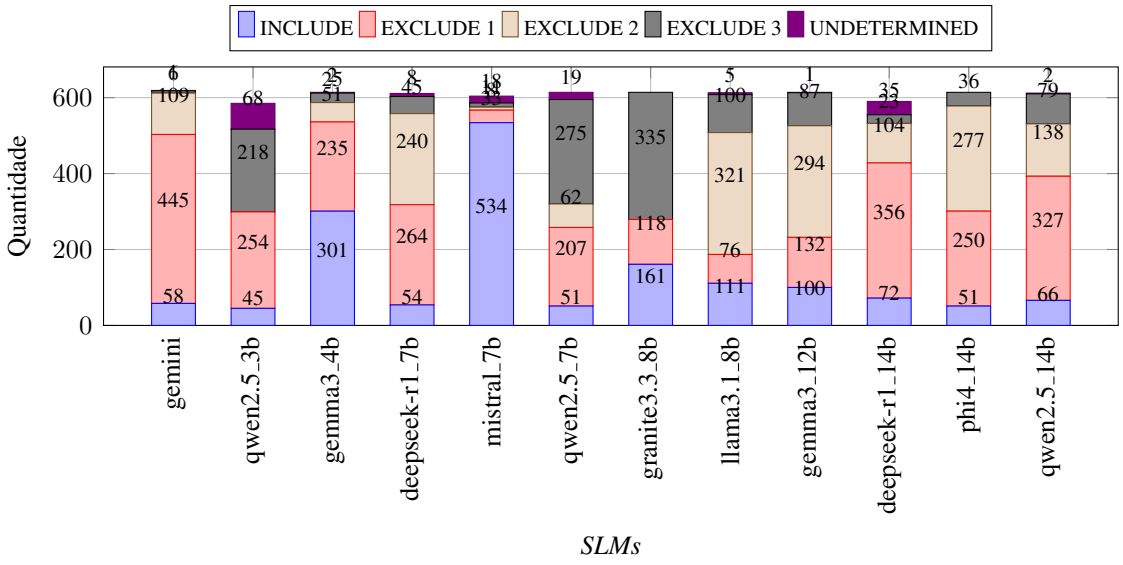

Figura 2. Distribuição das classificações por SLM (Abordagem 2)

Os *SLMs* com melhor performance foram *deepseek-r1_14b* marcando 64,01% e *qwen2.5_14b* com 59,45%, enquanto isso, *mistral_7b* e *granite3.3_8b* são os destaques negativos desta abordagem, com 14,5% e 25,9%, respectivamente. Portanto, a *Abordagem 2* apresenta resultados medianos em relação à concordância com o *baseline*, indicando que ainda é insuficiente.

4.3. Abordagem 3

Esta é a estratégia de *dividir para conquistar*, atribuindo tarefas mais simples (verificar apenas 1 critério de inclusão) para agentes paralelos, e um agente final que combina as avaliações dos anteriores. A utilização de um quarto agente (verificador) que conhece todo o contexto e julga as decisões dos três primeiros agentes aparentou ser uma via promissora a priori, porém os resultados obtidos demonstram que, na verdade, essa tarefa final tornou-se muito complexa para um *SLM*. Assim, para a avaliação dos primeiros 3 agentes, implementamos uma função *booleana* que combina as respostas deles conforme as regras apresentadas na Seção 3 e confrontamos com a decisão da classificação final do agente 4.

A Figura 4 apresenta as classificações dos agentes 1, 2 e 3 + Lógica Booleana em comparação com a decisão do agente 4 (verificador). Perceba no gráfico que, com

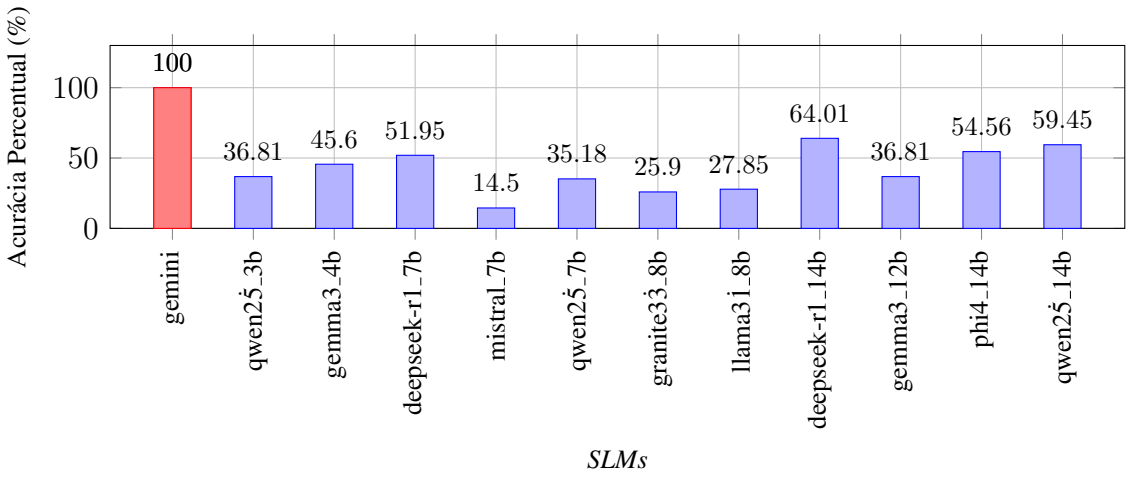

Figura 3. Percentual de concordância de cada *SLM* em relação ao *Gemini*

exceção do *baseline*, para cada SLM há duas barras, a da esquerda corresponde à resposta dos 3 agentes combinadas com a função booleana, e a da direita (hachurada) com as classificações do agente 4. O *mistral_7b* é um exemplo claro de perda de desempenho, onde os três primeiros agentes incluem apenas 50 artigos e depois do agente 4 a quantidade de inclusão sobe para 297, algo semelhante ocorre com o qwen2.5_3b.

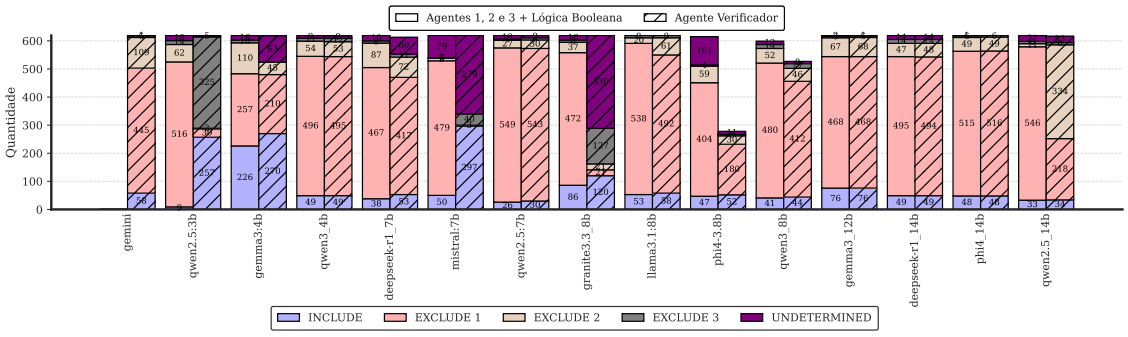

Figura 4. Distribuição das classificações feitas pelos agentes.

Também percebemos uma tendência dos modelos maiores manterem constante a quantidade de inclusão mesmo com o agente 4. Modelos médios têm tendência a aumentar os artigos incluídos a cada novo agente, o que, em geral, diminui a performance (acurácia em relação ao baseline). Os modelos menores também seguem uma tendência de aumento ao utilizar um novo agente. Os destaques são para o *Gemma 3:4b* que, desde o agente 3, já incluiu muito mais que os demais. No agente verificador, o maior aumento foi no *Qwen 2.5:3b*. Um destaque positivo é para o *Qwen 3:4b* com sua constância.

Para estudar os efeitos dessas distribuições, avaliamos a precisão, *recall* e *F1-score*, conforme apresentado nas Figuras 5 e 6 . A precisão significa: dos artigos que o *SLM* incluiu, qual a porcentagem de incluídos corretamente? Modelos com alta precisão irão retornar poucas inclusões irrelevantes, mesmo que muita coisa relevante fique de fora. O *recall* significa: entre os artigos que deveriam ser incluídos, quantos por cento o *SLM* incluiu? Ou seja, um modelo que fez muitas inclusões (inclusive irrelevantes)

pode apresentar um alto *recall*. Dessa forma, como o *F1-score* é a média harmônica entre precisão e *recall*, apresenta uma visão geral do desempenho do sistema, sendo uma métrica mais importante para esta tarefa.

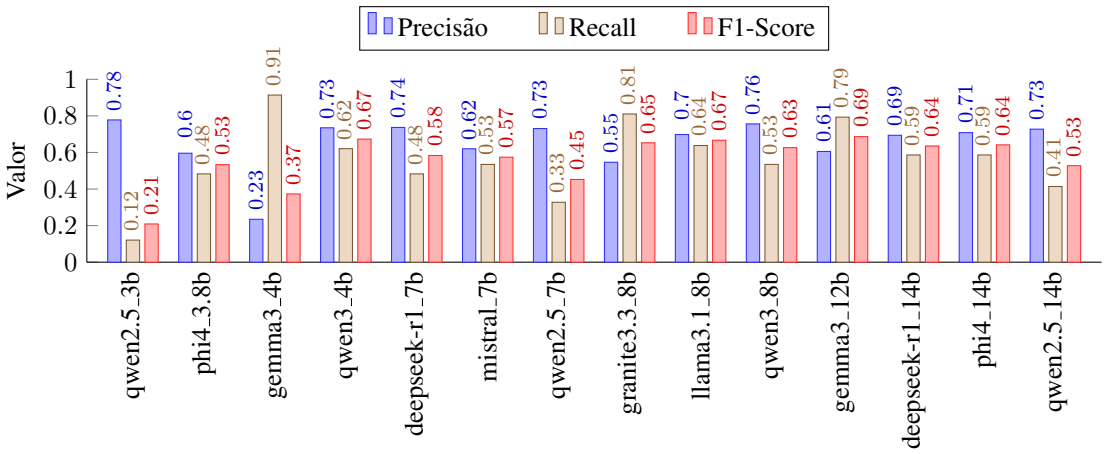

Figura 5. Métricas de desempenho dos *SLMs* utilizando agentes 1, 2, e 3 + lógica booleana.

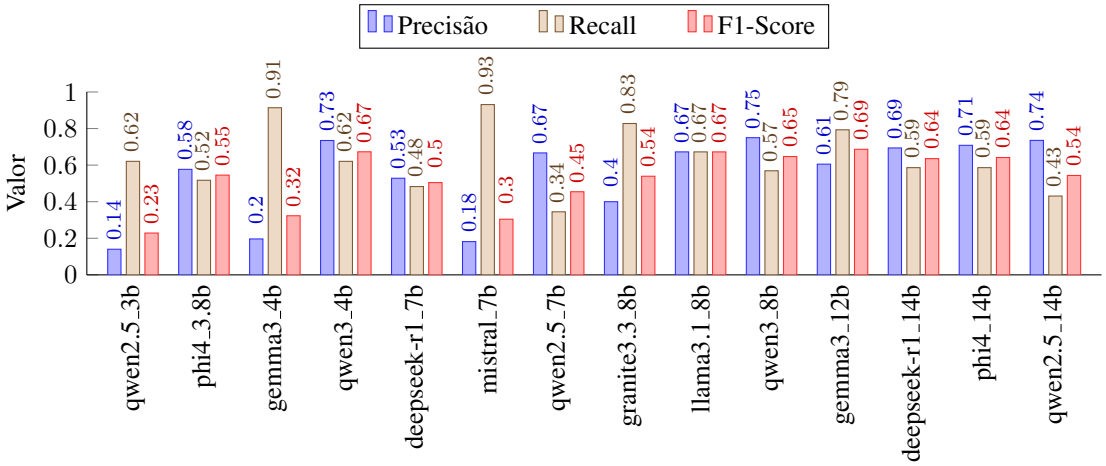

Figura 6. Métricas de desempenho dos *SLMs* utilizando o agente verificador.

Em relação à precisão, em geral, os *SLMs* tendem a manter a métrica constante ou diminuir ao se utilizar o quarto agente. O *Qwen2.5_3b* se destaca negativamente, diminuindo 64% de sua precisão ao utilizar o agente verificador, *mistral_7b* também apresenta uma queda considerável de 62% para 18%. Isso indica que aumentar a complexidade da tarefa geralmente piora o desempenho dos *SLMs*. Em relação ao *recall*, artigos que deveriam e realmente foram incluídos, alguns modelos como *Qwen 2.5:3b* demonstram um aumento notável, isso se dá pelo fato de ele ter incluído demais ao utilizar um novo agente, como mostra a Figura 4. Outro caso interessante é *Mistral:7b*, o qual o *recall* aumenta quando a precisão diminui, indicando novamente que utilizar um agente a mais, neste modelo, tende a piorar o desempenho do sistema. Os *F1-Scores* tendem a se manter constante ou diminuir ao adicionar o agente verificador. E de forma comparativa os melhores F1-Scores são dos modelos gemma3_12b, llama3.1_8b e qwen3_4b. A Figura 7 apresenta o percentual de concordância (acurácia) entre as decisões dos *SLMs* e o *baseline*.

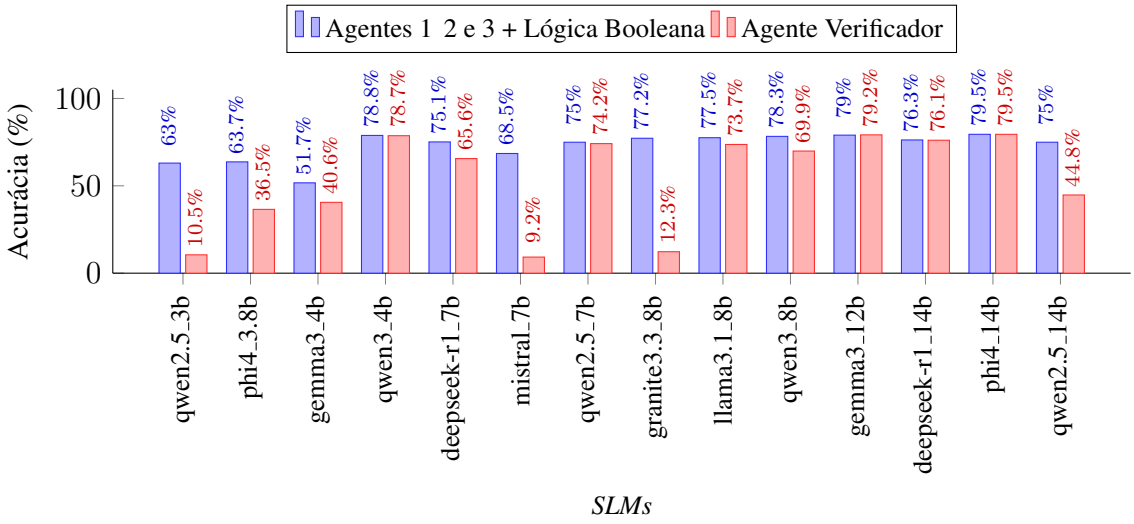

Figura 7. Concordância entre os *SLMs* e o *Gemini* na *Abordagem 2*.

Os modelos maiores (Deepseek:14b e Gemma:12b) mantiveram sua performance, exceto o Qwen 2.5:14b que teve uma piora significativa. Os modelos de tamanho intermediário (7b e 8b) tiveram redução quando a triagem avança para o próximo agente. Sendo os destaques para as maiores reduções: *Granite:8b* (redução 65 pontos percentuais), *Mistral:7b* (59,9%) e *Qwen 2.5:7b* (52,9%). Com exceção do *Qwen 3:4b*, os menores modelos (3b e 4b) já possuem menor desempenho no agente 3 e também seguem a tendência de redução quando são enviados ao agente verificador. O resultado mais surpreendente é o do *Qwen 3:4b* com uma performance idêntica à dos maiores modelos, ou seja, desempenho próximo de 80% desde o agente 3 e se mantém passando pelo agente verificador.

5. Conclusões e pesquisas futuras

Este trabalho apresentou 3 abordagens para triagem de trabalhos científicos usando SLMs, nas quais os critérios de exclusão são definidos pelo pesquisador, conforme seu protocolo de RSL. Apesar da *Abordagem 1* apresentar acurácias elevadas, ela é insuficiente, pois não define o motivo exato da exclusão, afetando diretamente a rastreabilidade do processo de RSL. Os *SLMs* não apresentaram baixo desempenho, utilizando a *Abordagem 2*, em relação ao *baseline*. A *Abordagem 3* foi implementada em 2 versões, apresentando bom desempenho e rastreabilidade, utilizando multiagentes com uma lógica booleana.

Os resultados obtidos apontam que *SLMs* conseguem ter bom desempenho na tarefa de triagem de artigos, principalmente quando recebem tarefas estratificadas mais simples segundo uma estratégia do tipo *dividir para conquistar*, relativamente se aproximando ao resultado do modelo utilizado como *baseline*, o *Gemini 2.0-flash*. Dentre os *SLMs* testados, o *Qwen 3:4b* se destaca positivamente por ser pequeno, oferecer custo-benefício de inferência e apresentar uma acurácia de 78.8% em relação ao *baseline*.

Trabalhos futuros podem combinar o que têm de melhor nas abordagens 1 e 3 em uma solução mista, aproveitando a alta taxa de acurácia da *Abordagem 1* e a definição dos critérios da *Abordagem 3*. Trabalhos futuros também podem explorar a combinação de agentes de diferentes tamanhos, usando *SLMs* menores para tarefas mais simples e mai-

ores para as mais complexas. Os erros de formatação de alguns modelos apontam para a utilização de um agente revisor para garantir a formatação correta. Além disso, é importante aplicar a mesma metodologia deste trabalho em outras RSLs para confirmar que os resultados não estão enviesados pelo tema. A integração e a sistematização da etapa de triagem abordada neste trabalho com as outras fases do processo de revisão serão realizadas para a construção de um assistente inteligente de RSL. Portanto, o aprendizado adquirido nesta experimentação deverá ser utilizado para a implementação das outras etapas da RSL.

6. Agradecimentos

Agradecimentos para o Programa de Pós-graduação em Ciência da Computação UERN/UFERSA e ao laboratório GESyCA da UFERSA. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

Referências

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., and Molchanov, P. (2025). Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*.

Borah, R., Brown, A. W., Capers, P. L., and Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ Open*, 7(2).

Colangelo, M. T., Guizzardi, S., Meleti, M., Calciolari, E., and Galli, C. (2025). Performance comparison of large language models for efficient literature screening. *BioMedInformatics*, 5(2).

Delgado-Chaves, F. M., Jennings, M. J., Atalaia, A., Wolff, J., Horvath, R., Mamdouh, Z. M., Baumbach, J., and Baumbach, L. (2025). Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences*, 122(2):e2411962122.

Galli, C., Gavrilova, A. V., and Calciolari, E. (2025). Large language models in systematic review screening: Opportunities, challenges, and methodological considerations. *Information*, 16(5).

Gusenbauer, M. and Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. *Research Synthesis Methods*, 11(2):181–217.

Haddaway, N. R. and Westgate, M. J. (2018). Predicting the time needed for environmental systematic reviews and systematic maps. *Conservation Biology*, 33(2):434–443.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K.,

Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Keele University and University of Durham, UK.

LangChain Team and Contributors (2025). Langchain.

Li, M., Sun, J., and Tan, X. (2024). Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Systematic Reviews*, 13:219.

Luo, H., Liu, P., and Esping, S. (2023). Exploring small language models with prompt-learning paradigm for efficient domain-specific text classification. *arXiv preprint arXiv:2309.14779*.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J. B., Rome, A., Shi, A., and Oak, S. (2025). The ai index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

Mechanism, S. A. (2024). Successful and timely uptake of artificial intelligence in science in the eu: evidence review report.

Mulrow, C. D. (1994). Systematic reviews: Rationale for systematic reviews. *BMJ*, 309(6954):597–599.

Ollama Team and Contributors (2025). Ollama.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Sampson, M., Tetzlaff, J., and Urquhart, C. (2011). Precision of healthcare systematic review searches in a cross-sectional sample. *Research Synthesis Methods*, 2(2):119–125.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity.

Telenti, A., Auli, M., Hie, B. L., Maher, C., Saria, S., and Ioannidis, J. P. A. (2024). Large language models for science and medicine. *European Journal of Clinical Investigation*, 54(6):e14183. e14183 EJCI-2023-1951.R1.